Learning Frame Similarity using Siamese networks for Audio-to-Score Alignment

Ruchit Agrawal
Centre for Digital Music
Queen Mary University of London
London, United Kingdom
r.r.agrawal@qmul.ac.uk

Simon Dixon
Centre for Digital Music
Queen Mary University of London
London, United Kingdom
s.e.dixon@qmul.ac.uk

***Abstract*—Audio-to-score alignment aims at generating an accurate mapping between a performance audio and the score of a given piece. Standard alignment methods are based on Dynamic Time Warping (DTW) and employ handcrafted features, which cannot be adapted to different acoustic conditions. We propose a method to overcome this limitation using learned frame similarity for audio-to-score alignment. We focus on offline audio-to-score alignment of piano music. Experiments on music data from different acoustic conditions demonstrate that our method achieves higher alignment accuracy than a standard DTW-based method that uses handcrafted features, and generates robust alignments whilst being adaptable to different domains at the same time.**

***Index Terms*—Music Information Retrieval, Audio-to-Score Alignment, Siamese networks, Convolutional Neural Networks, Dynamic Time Warping**

I. Introduction

The significance of neural networks for signal processing was pointed out early by [1], [2], and their efficacy for Music Information Retrieval (MIR) has been demonstrated for a variety of tasks like music generation [3], music transcription [4] as well as music alignment [5]. Audio-to-score alignment is the task of finding the optimal mapping between a performance and the score for a given piece of music. Dynamic Time Warping (DTW) [6] has been the de facto standard for this task, typically incorporating handcrafted features [7]–[9]. The primary limitation of handcrafted features lies in their inability to adapt to different acoustic settings and thereby model real world data in a robust manner, in addition to not being optimized for the task at hand.

This paper presents a novel method for DTW-based audio-to-score alignment, which does not depend on handcrafted features, but learns them directly from the music data at the frame level. We propose learning a frame similarity matrix using neural networks which is then passed on to a DTW algorithm that computes the optimal warping path through the matrix, yielding the alignment. We propose the use of twin Siamese networks [10] each containing a Convolutional Neural Network (CNN) [11] architecture for learning frame similarity. The advantage of our method is that it is is efficiently able to learn meaningful representations for DTW directly from data and is thereby adaptable to different acoustic settings.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068.

We conduct experiments on piano music using our approach and test its performance on the Mazurka dataset [12], which contains recordings from different eras spanning various acoustic conditions; and demonstrate improvements over *MATCH* [13], a standard DTW-based method that uses handcrafted features. We additionally explore two methods to improve the performance of our baseline models, namely salience representations [14] and data augmentation.

To the authors' knowledge, this is the first method to employ learned frame similarities using Siamese CNNs for audio-to-score alignment. Additionally, this is the first method to incorporate pitch salience for audio-to-score alignment to the authors' knowledge. The rest of the paper is organized as follows: We describe prior work and our relation to it in Section II. Section III details our proposed method and model pipeline. The experimentation conducted and results obtained using our method are described in Section IV. We present the conclusions of the present research and highlight possible directions for future work in Section V.

II. Related Work

Early works on feature learning for Music Information Retrieval (MIR) employ algorithms like Conditional Random Fields [15] or deep belief networks [16], whereas recent work in this direction is moving towards the usage of deep neural networks [17]. Work specifically on learning features for audio-to-score alignment has focused on the evaluation of current feature representations [18], learning features for alignment using a Multi Layer Perceptron [19], and learning a mapping several common audio representations based on a best-fit criterion [20]. Recently, transposition-invariant features were proposed for music alignment [21], however these features while being robust to transposition, are sensitive to large tempo variations and underperform in such situations. [22] is a recent work on score following, a task related to audio-to-score alignment. While they employ reinforcement learning to train a score follower in real time, we focus on robust offline alignment across various acoustic conditions using frame similarity learning.

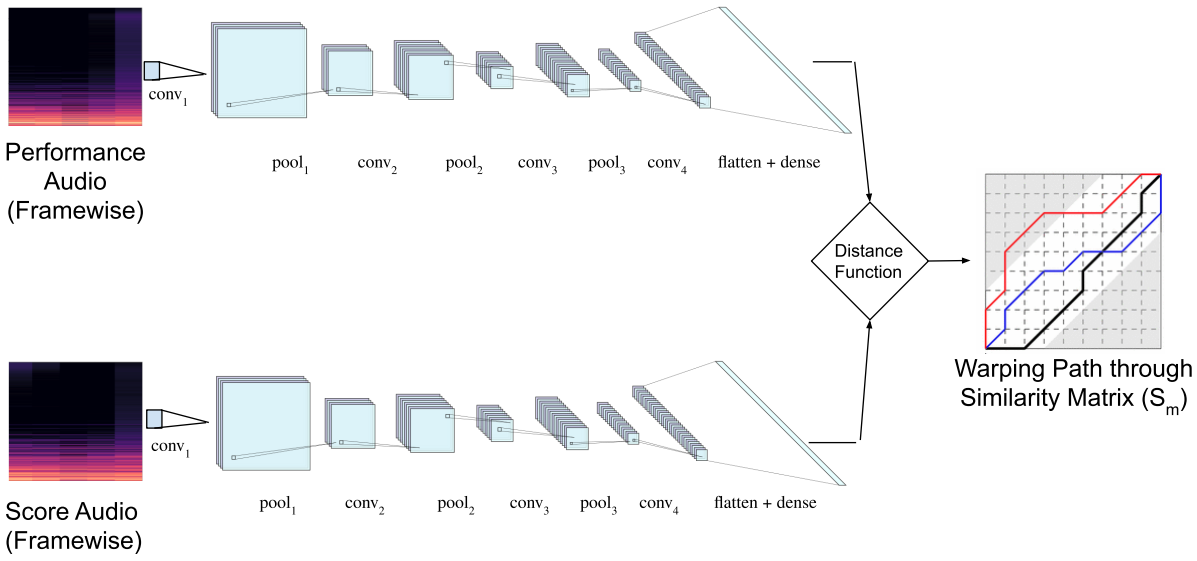

Fig. 1: Model Pipeline
$conv_i$: i_{th} *convolution layer* $pool_j$: j_{th} *pooling layer*
$flatten$: *Flatten layer* $dense$: *Fully connected layer*

Another direction which sets the context for our work is sound similarity; approaches to which include capturing music segment similarity using two-dimensional Fourier-magnitude coefficients [23], similarity network fusion to combine different frame-level features for hierarchy identification of repeated sections in music [24], and application of Siamese Neural Networks for content-based audio retrieval [25]. The closest work to ours which employs the notion of learned sound similarity for music alignment is [19], to the authors' knowledge. While they use a Multi-Layer Perceptron to compute if two frames are the same or not, we compute frame similarity using Siamese CNNs. In addition to using an enhanced framework which is suitable for the similarity detection task, our work differs from them in that we also compute the extent of similarity in the form of non-binary distances and use this distance (or dissimilarity) matrix further for alignment. We additionally employ deep salience representations, which prove to be an effective method to improve alignment accuracy over our baseline models.

III. Proposed Methodology

We propose a novel method for DTW-based audio-to-score alignment that uses Siamese neural networks. We additionally employ deep salience representations [14] to improve model performance in data-scarce conditions. We describe the method in detail in the subsequent subsections.

A. Siamese Convolutional Neural Networks

The standard feature representation choice for music alignment is a time-chroma representation [26] generated from the log-frequency spectrogram, which is not trainable on real data, and thereby not adaptable to different acoustic settings. We override the feature engineering step and focus on learning frame similarity using Convolutional Neural Networks (CNNs), since they can jointly optimize the representation of input data conditioned on the similarity measure being used. We employ a Siamese Convolutional Neural Network, a class of neural network architectures that contains two or more identical subnetworks [10] for this task.

We train a Siamese CNN, akin to that prototyped in [27], to compute a frame similarity matrix S_m to be fed to DTW to generate alignment. Figure 1 gives an overview of our model pipeline. In order to keep the modality constant, we first convert the MIDI files to audio through FluidSynth [28] using piano soundfonts. The two audio inputs are converted to a low-level spectral representation using a Short Time Fourier Transform, with a hop size of 23 ms and a hamming window of size 46 ms. Our training data contains synchronized audio and MIDI files, so it is straightforward to extract matching frame pairs. For each matching pair, we randomly select a non-matching pair (using MIDI-information) in order to have a balanced training set. The inputs to the Siamese network are labelled frame pairs from the performance audio and the synthesized MIDI respectively. We employ the contrastive loss function [29] while training our models. We choose this formulation over a standard classification loss function like cross entropy since our objective is to differentiate between two audio frames. Let $X = (X_1, X_2)$ be the pair of inputs X_1 and X_2, W be the set of parameters to be learnt and Y be the target binary label (Y = 0 if they match and 1 if otherwise). Task-specific loss functions have shown promising results in the fields of image processing and natural language

processing [30], [31]. The contrastive loss function for each tuple is computed as follows:

$$L(W,X,Y) = (1-Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{max(0, m - D_W)\}^2 \quad (1)$$

where m is the margin for dissimilarity and D_W is the Euclidean Distance between the outputs of the subnetworks. Pairs with dissimilarity greater than m do not contribute to the loss function. More formally, D_W can be expressed as follows:

$$D_W(X) = \sqrt{\{G_W(X_1) - G_W(X_2)\}^2} \quad (2)$$

where G_W is the output of each twin subnetwork for the inputs X_1 and X_2. Since it is a distance-based loss, it tries to ensure that semantically similar examples are embedded close to each other, which is a desirable trait for extracting alignments.

The Siamese network thus learns to classify the sample pairs as similar or dissimilar. This is done for each audio frame pair and the similarity matrix thus generated is then passed on to a DTW-based algorithm to generate the alignment path. DTW generates an alignment between two sequences A= $(a_1, a_2, ..., a_m)$ and B = $(b_1, b_2, ..., b_n)$ by comparing them using a local cost function, at each point, with the goal of minimizing the overall cost. The path which yields this minimum overall cost is then the optimal alignment between the two sequences. Formally, it can be represented as follows:

$$D(i,j) = d(i,j) + min \begin{cases} D(i, j-1) \\ D(i-1, j) \\ D(i-1, j-1) \end{cases} \quad (3)$$

where $d(i,j)$ is the distance measure (local cost) between points a_i and b_j; and $D(i,j)$ is the total cost for the path which generates the optimal alignment between the sequences $A_{1..i}$ and $B_{1..j}$. We employ Euclidean distance as our distance measure and the DTW framework from [32] to compute the warping paths.

B. Deep Salience Representations

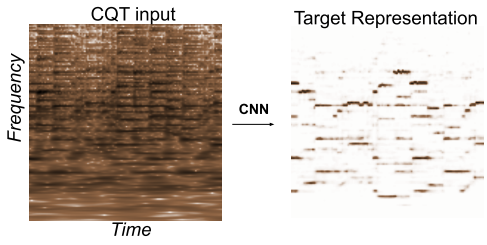

Fig. 2: Salience representations to address data sparsity

We employ deep salience representations [14] for effective training of our models. These are time-frequency representations aimed at estimating the likelihood of a pitch being present in the audio. Figure 2 shows an example of a salience representation.

The primary motivation behind using such a representation is that it de-emphasizes non-pitched content and emphasizes harmonic content, thereby aiding training in data-scarce conditions. We employ the model proposed by [14], trained to learn a series of convolutional filters, constraining the target salience representation to have values between 0 and 1, with larger values corresponding to time-frequency bins where fundamental frequencies are present. The model is trained to minimize the cross entropy loss as follows:

$$L(y, \hat{y}) = -ylog(\hat{y}) - (1-y)log(1-\hat{y}) \quad (4)$$

where both y and $\hat{y}$ are continuous values between 0 and 1.

We compare the performance obtained using salience representations with that obtained using the Short-Time Fourier Transform (STFT) and Constant-Q Transform (CQT) of the raw audios. We employ these input representations for comparative purposes. We employ a hop size of 23 ms and a hamming window of size 46 ms. We employ a CQT with 24 bins per octave, with the first bin corresponding to frequency 65.4 Hz (midi note C2).

IV. Experiments and Results

A. Experimental Setup

We employ the MAPS database [33], the Saarland database [34] and the Mazurka dataset [12] for our experiments. From the original MAPS database, which contains synthesized MIDI-aligned audio for a range of acoustic settings, we select the subset *MUS* containing complete pieces of piano music, and append it to the Saarland database. We split the resultant database comprising 288 recordings randomly into sets of 230 and 58 recordings. These sets form our training and validation sets respectively. We test the performance of our models on the Mazurka dataset [12], which contains recordings of Chopin's Mazurkas dating from 1902 to the early 2000s, thereby spanning across various acoustic settings. This dataset contains annotations of beat times for five Mazurka pieces. The alignment error for these pieces has a standard deviation of 11 ms.

TABLE I: Architecture of our model

Type of layer	Input size	Kernels	Kernel size
Convolution	$128 * 128 * 3$	64	$5 * 5$
Max-Pooling	$128 * 128 * 64$	1	$2 * 2$
Convolution	$64 * 64 * 64$	128	$5 * 5$
Max-Pooling	$64 * 64 * 128$	1	$2 * 2$
Convolution	$32 * 32 * 128$	256	$3 * 3$
Max-Pooling	$32 * 32 * 256$	1	$2 * 2$
Convolution	$16 * 16 * 256$	512	$3 * 3$
Flatten	$16 * 16 * 512$	-	-
Fully Connected	131072	-	-

Our Siamese model has four convolutional layers of varying dimensionality followed by a fully connected layer to generate the similarity output. The outputs of each layer are passed through rectified linear units in order to add non-linearity,

TABLE II: Results of our models

Model	*Binary Matrix*				*Distance Matrix*			
	<25ms	**<50ms**	**<100ms**	**<200ms**	**<25ms**	**<50ms**	**<100ms**	**<200ms**
$MATCH$ [13]	-	-	-	-	64.8	72.1	77.6	83.7
DTW_{Chroma}	-	-	-	-	62.9	70.5	76.3	82.4
$MLP_{Semigram}$ [19]	63.8	69.5	77.2	83.4	-	-	-	-
$SCNN_{STFT}$	65.6	71.9	78.1	84.8	67.2	73.4	78.7	85.6
$SCNN_{CQT}$	66.4	73.1	78.7	85.3	68.1	74.8	80.1	86.7
$SCNN_{Chroma}$	67.1	74.6	79.2	86.1	69.4	75.1	80.7	87.2
$SCNN_{Sal}$	68.2	75.3	**81.4**	**87.8**	70.3	76.7	82.1	88.4
$SCNN_{CQT+DA}$	67.9	74.4	80.8	86.7	69.6	75.4	81.6	87.9
$SCNN_{Sal+DA}$	**69.4**	**76.4**	81.2	87.5	**71.7**	**78.2**	**83.3**	**90.1**

followed by batch normalization before being passed as inputs to the next layer. The detailed architecture of our model is given in Table I.

We conduct experiments using two different mechanisms for computing the similarity matrix S_m:

- Using binary labels: We directly employ the outputs of the Siamese CNN, whereby 0 and 1 correspond to similar and dissimilar pairs respectively.
- Using distances: We employ the distance D_W computed as part of the loss, which directly corresponds to the dissimilarity between the two inputs.

We generate an alignment path through this matrix using DTW, through a readily available implementation in Python [32]. For our Siamese models trained without data augmentation, the naming convention we employ is $SCNN_x$, where x is the feature representation used during training. We also report results obtained using data augmentation. We generate 20% additional training samples by employing a random pitch shift of up to ± 30 cents, using librosa [35]. These models are named $SCNN_{CQT+DA}$ and $SCNN_{Sal+DA}$ for the CQT and the salience representations respectively.

B. Results and Discussion

We compare the performance of our models with *MATCH* [13]; a DTW algorithm using Chroma features [26]; and the Multi-Layer Perceptron Model proposed by [19] ($MLP_{Semigram}$). We compute the error e_i = $t_i{}^e$ - $t_i{}^r$, defined as the time difference between the alignment positions of corresponding events in the reference $t_i{}^r$ and the estimated alignment time $t_i{}^e$ for score event i. We show results for accuracy in percentage for events which are aligned within an error of up to 25 ms, 50ms, 100ms and 200ms respectively. The results obtained by our models are given in Table II.

Our models outperform DTW-based algorithms that employ handcrafted features as well as an MLP framework which learns binary similarity labels (Table II, rows 1-5). The CQT representation ($SCNN_{CQT}$) yields better results than the STFT representation ($SCNN_{STFT}$), we argue that this is due to the nature of the CQT, which is a more musically meaningful representation. Our Siamese model trained using the Chroma representation ($SCNN_{Chroma}$) outperforms the DTW-based method using the same representation (DTW_{Chroma}), suggesting that frame similarity learnt from real data is effective at generating robust alignment. Additionally, we observe the trend that the models trained using a non-binary distance matrix outperform those trained on binary matrices (Table II, columns 6-9). We speculate that thresholding the similarity into binary labels discards potentially useful information and the distances facilitate the DTW algorithm to take better long-term decisions. Both salience representations ($SCNN_{Sal}$) and data augmentation ($SCNN_{DA}$) prove to be effective to improve the performance of our model over $SCNN_{CQT}$, with salience representations contributing to greater improvements. We posit that using salience representations makes it easier for the model to learn meaningful features from the input representations, since it emphasizes pitched content. Improvements using data augmentation can be attributed to the fact that pianos are not always tuned to $A = 440Hz$ in the real world, and often the relative intervals are also not tuned perfectly, hence comparison with MIDI files in such cases might lead to false negatives. Data augmentation ensures that the disparity between our training and test conditions is minimized by simulating more real-world like conditions in our training data. A combination of distance matrix, salience representations and data augmentation yields the best results ($SCNN_{Sal+DA}$), as can be seen from Table II, row 8, columns 6-9.

Our results demonstrate that frame similarity learning using Siamese neural networks is a promising method for audio-to-score alignment. The principal advantage of this approach over traditional feature choices (like chroma features or MFCCs) is the ability to learn directly from data, which provides higher relevance and adaptability. Both the Siamese network and the pitch salience network are trainable, and thereby adaptable to real world conditions. We plan to explore domain adaptation of our models in the future. A limitation of our method is that it cannot handle structural changes, since DTW generates a monotonically increasing warping path. This could potentially be mitigated by employing an enhanced DTW framework like jump-DTW [36] alongside our Siamese model.

V. Conclusion and Future Work

We presented a novel method for offline audio-to-score alignment using learned similarities via a Siamese convolutional network architecture. We demonstrated that our approach is capable of generating robust alignments for piano music across various acoustic conditions. Our models outperform traditional methods based on Dynamic Time Warping that rely on handcrafted features, as well as a Multi Layer Perceptron model which learns binary similarity between audio frames. We also demonstrated that salience representations and data augmentation are effective techniques to improve alignment accuracy. In the future we plan to incorporate attention into the convolutional models to aid training and improve performance. We would also like to explore other model architectures and work on learning the features as well as the alignments in a completely end-to-end manner.

References

[1] Jenq-Nen Hwang, Sun-Yan Kung, Mahesan Niranjan, and Jose C Principe, "The past, present, and future of neural networks for signal processing," *IEEE Signal Processing Magazine*, vol. 14, no. 6, pp. 28–48, 1997.

[2] Fa-Long Luo and Rolf Unbehauen, "Applied neural networks for signal processing," 1999.

[3] Douglas Eck and Juergen Schmidhuber, "A first look at music composition using lstm recurrent neural networks," 2002.

[4] Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck, "Onsets and frames: Dual-objective piano transcription," 2018.

[5] Matthias Dorfer, Jan Hajič Jr, Andreas Arzt, Harald Frostel, and Gerhard Widmer, "Learning audio–sheet music correspondences for cross-modal retrieval and piece identification," *Transactions of the International Society for Music Information Retrieval*, vol. 1, no. 1, 2018.

[6] Hiroaki Sakoe and Seibi Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[7] Simon Dixon, "An on-line time warping algorithm for tracking musical performances.," in *International Joint Conference on Artificial Intelligence*, 2005, pp. 1727–1728.

[8] Sebastian Ewert, Meinard Muller, and Peter Grosche, "High resolution audio synchronization using chroma onset features," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 1869–1872.

[9] Andreas Arzt, Gerhard Widmer, and Simon Dixon, "Adaptive distance normalization for real-time music tracking," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 2689–2693.

[10] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah, "Signature verification using a siamese time delay neural network," in *Advances in Neural Information Processing Systems*, 1994, pp. 737–744.

[11] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio, "Object recognition with gradient-based learning," in *Shape, Contour and Grouping in Computer Vision*, pp. 319–345. Springer, 1999.

[12] Craig Stuart Sapp, "Comparative analysis of multiple musical performances.," in *International Society for Music Information Retrieval (ISMIR)*, 2007, pp. 497–500.

[13] Simon Dixon and Gerhard Widmer, "Match: A music alignment tool chest.," in *International Society for Music Information Retrieval*, 2005, pp. 492–497.

[14] Rachel M Bittner, Brian McFee, Justin Salamon, Peter Li, and Juan Pablo Bello, "Deep salience representations for f0 estimation in polyphonic music.," in *International Society for Music Information Retrieval (ISMIR)*, 2017, pp. 63–70.

[15] Cyril Joder, Slim Essid, and Gael Richard, "Learning optimal features for polyphonic audio-to-score alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2118–2128, 2013.

[16] Erik M Schmidt, Jeffrey J Scott, and Youngmoo E Kim, "Feature learning in dynamic environments: Modeling the acoustic structure of musical emotion.," in *International Society for Music Information Retrieval (ISMIR)*, 2012, pp. 325–330.

[17] John Thickstun, Zaid Harchaoui, and Sham Kakade, "Learning features of music from scratch," *arXiv preprint arXiv:1611.09827*, 2016.

[18] Cyril Joder, Slim Essid, and Gaël Richard, "A comparative study of tonal acoustic features for a symbolic level music-to-score alignment," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 409–412.

[19] Özgür İzmirli and Roger B Dannenberg, "Understanding features and distance functions for music sequence alignment.," in *International Society for Music Information Retrieval (ISMIR)*, 2010, pp. 411–416.

[20] Cyril Joder, Slim Essid, and Gaël Richard, "Optimizing the mapping from a symbolic to an audio representation for music-to-score alignment," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 121–124.

[21] Andreas Arzt and Stefan Lattner, "Audio-to-score alignment using transposition-invariant features," 2018.

[22] Matthias Dorfer, Florian Henkel, and Gerhard Widmer, "Learning to listen, read, and follow: Score following as a reinforcement learning game," in *International Society for Music Information Retrieval*, 2018.

[23] Oriol Nieto and Juan Pablo Bello, "Music segment similarity using 2d-fourier magnitude coefficients," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 664–668.

[24] Christopher J Tralie and Brian McFee, "Enhanced hierarchical music structure annotations via feature level similarity fusion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 201–205.

[25] Pranay Manocha, Rohan Badlani, Anurag Kumar, Ankit Shah, Benjamin Elizalde, and Bhiksha Raj, "Content-based representations of audio using siamese neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 3136–3140.

[26] Mark A Bartsch and Gregory H Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 96–104, 2005.

[27] Ruchit Agrawal and Simon Dixon, "A hybrid approach to audio-to-score alignment," .

[28] David Henningsson and FluidSynth Developer Team, "Fluidsynth real-time and thread safety challenges," in *Proceedings of the 9th International Linux Audio Conference, Maynooth University, Ireland*, 2011, pp. 123–128.

[29] Raia Hadsell, Sumit Chopra, and Yann LeCun, "Dimensionality reduction by learning an invariant mapping," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2006, vol. 2, pp. 1735–1742.

[30] Ce Qi and Fei Su, "Contrastive-center loss for deep neural networks," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 2851–2855.

[31] Tebbifakhr Amirhossein, Ruchit Rajeshkumar Agrawal, Rajen Chatterjee, Matteo Negri, and Marco Turchi, "Multi-source transformer with combined losses for automatic post editing," in *Third Conference on Machine Translation (WMT)*. The Association for Computational Linguistics, 2018, pp. 859–865.

[32] Toni Giorgino et al., "Computing and visualizing dynamic time warping alignments in r: The dtw package," *Journal of Statistical Software*, vol. 31, no. 7, pp. 1–24, 2009.

[33] Valentin Emiya, Roland Badeau, and Bertrand David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2009.

[34] Meinard Müller, Verena Konz, Wolfgang Bogler, and Vlora Arifi-Müller, "Saarland music data (smd)," in *International Society for Music Information Retrieval: late breaking session*, 2011.

[35] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*, 2015, pp. 18–25.

[36] Christian Fremerey, Meinard Müller, and Michael Clausen, "Handling repeats and jumps in score-performance synchronization.," in *International Society for Music Information Retrieval (ISMIR)*, 2010, pp. 243–248.